Fitting Regression Models

In general, suppose that there is a single **dependent variable** or **response** y that depends on k **independent** or **regressor variables**, for example, $x_1, x_2, \ldots, x_k$. The relationship between these variables is characterized by a mathematical model called a **regression model**. The regression model is fit to a set of sample data. In some instances, the experimenter knows the exact form of the true functional relationship between y and $x_1, x_2, \ldots, x_k$, say $y = \phi(x_1, x_2, \ldots, x_k)$. However, in most cases, the true functional relationship is unknown, and the experimenter chooses an appropriate function to approximate ϕ. Low-order polynomial models are widely used as approximating functions.

Linear Regression models

- Suppose that we wish to develop an empirical model relating the viscosity of a polymer to the temperature and the catalyst feed rate.

- A model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad (10.1)$$

where y represents the viscosity, x_1 represents the temperature, and x_2 represents the catalyst feed rate. This is a **multiple linear regression model** with two independent variables. We often call the independent variables **predictor variables** or **regressors**. The term **linear** is used because Equation 10.1 is a linear function of the unknown parameters β_0, β_1, and β_2. The model describes a plane in the two-dimensional x_1, x_2 space. The parameter β_0 defines the intercept of the plane. We sometimes call β_1 and β_2 *partial regression coefficients* because β_1 measures the expected change in y per unit change in x_1 when x_2 is held constant and β_2 measures the expected change in y per unit change in x_2 when x_1 is held constant.

In general, the response variable y may be related to k regressor variables. The model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad \textbf{(10.2)}$$

is called a *multiple linear regression model* with k regressor variables. The parameters β_j, $j = 0, 1, \ldots, k$, are called the **regression coefficients**. This model describes a hyperplane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response y per unit change in x_j when all the remaining independent variables x_i $(i \neq j)$ are held constant.

Models that are more complex in appearance than Equation 10.2 may often still be analyzed by multiple linear regression techniques. For example, consider adding an interaction term to the first-order model in two variables, say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \qquad \textbf{(10.3)}$$

If we let $x_3 = x_1 x_2$ and $\beta_3 = \beta_{12}$, then Equation 10.3 can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \qquad \textbf{(10.4)}$$

which is a standard multiple linear regression model with three regressors.

Estimation of the Parameters in Linear Regression Models

The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. Suppose that $n > k$ observations on the response variable are available, say $y_1, y_2, \ldots, y_n$. Along with each observed response y_i, we will have an observation on each regressor variable and let x_{ij} denote the ith observation or level of variable x_j. The data will appear as in Table 10.1. We assume that the error term ϵ in the model has $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$ and that the $\{\epsilon_i\}$ are uncorrelated random variables.

We may write the model equation (Equation 10.2) in terms of the observations in Table 10.1 as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \quad i = 1, 2, \ldots, n \qquad \textbf{(10.7)}$$

The method of least squares chooses the β's in Equation 10.7 so that the sum of the squares of the errors, ϵ_i, is minimized. The least squares function is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The function L is to be minimized with respect to $\beta_0, \beta_1, \ldots, \beta_k$.

■ TABLE 10.1
Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮		⋮
y_n	x_{n1}	x_{n2}	...	x_{nk}

- To estimate the parameters, we have to solve:

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
&\vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i \qquad \textbf{(10.10)}
\end{aligned}
$$

These equations are called the **least squares normal equations**. Note that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the least squares estimators of the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

Matrix representation

- Equation 10.7(mentioned in slide 5), may be written in matrix notation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In general, $\mathbf{y}$ is an $(n \times 1)$ vector of the observations, $\mathbf{X}$ is an $(n \times p)$ matrix of the levels of the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of random errors.

Estimate of β and variance

the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (10.13)$$

- To estimate variance, we consider the sum of square of residuals as:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

- It can be shown the sum of square can be written as :
- $$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} \qquad (10.17)$$

- The method of least squares produces an unbiased estimator of the parameter β in the linear regression model.
- By unbiased we mean:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

The variance property of $\hat{\boldsymbol{\beta}}$ is expressed in the **covariance matrix**:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad \textbf{(10.19)}$$

If σ^2 in Equation 10.19 is replaced with the estimate $\hat{\sigma}^2$ we obtain an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$. The square roots of the main diagonal elements of this matrix are the **standard errors** of the model parameters.

EXAMPLE 10.1

Sixteen observations on the viscosity of a polymer (y) and two process variables—reaction temperature (x_1) and catalyst feed rate (x_2)—are shown in Table 10.2. We will fit a multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

to these data.

■ TABLE 10.2
Viscosity Data for Example 10.1 (viscosity in centistokes @ 100°c)

Observation	Temperature (x_1, °C)	Catalyst Feed Rate (x_2, lb/h)	Viscosity
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

The **X** matrix and **y** vector are

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix} = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133{,}560 & 14{,}946 \\ 164 & 14{,}946 & 1{,}726 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix} = \begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}$$

The least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

or

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 14.176004 & -0.129746 & -0.223453 \\ -0.129746 & 1.429184 \times 10^{-3} & -4.763947 \times 10^{-5} \\ -0.223453 & -4{,}763947 \times 10^{-5} & 2.222381 \times 10^{-2} \end{bmatrix} \begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix} = \begin{bmatrix} 1566.07777 \\ 7.62129 \\ 8.58485 \end{bmatrix}$$

The least squares fit, with the regression coefficients reported to two decimal places, is

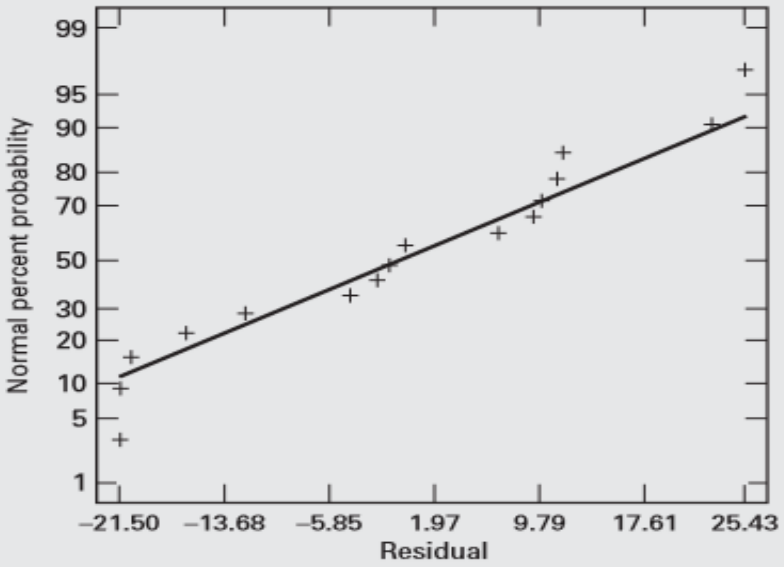
$$\hat{y} = 1566.08 + 7.62x_1 + 8.58x_2$$

The first three columns of Table 10.3 present the actual observations y_i, the predicted or fitted values $\hat{y}_i$, and the residuals. Figure 10.1 is a normal probability plot of the residuals. Plots of the residuals versus the predicted values $\hat{y}_i$ and versus the two variables x_1 and x_2 are shown in Figures 10.2, 10.3, and 10.4, respectively. Just as in designed experiments, residual plotting is an integral part of regression model building. These plots indicate that the variance of the observed viscosity tends to increase with the magnitude of viscosity. Figure 10.3 suggests that the variability in viscosity is increasing as temperature increases.

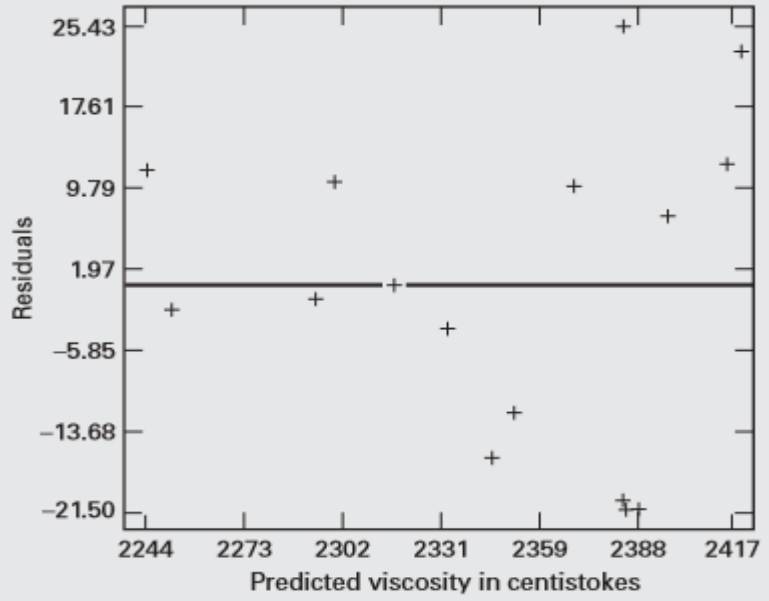
■ **TABLE 10.3**

Predicted Values, Residuals, and Other Diagnostics from Example 10.1

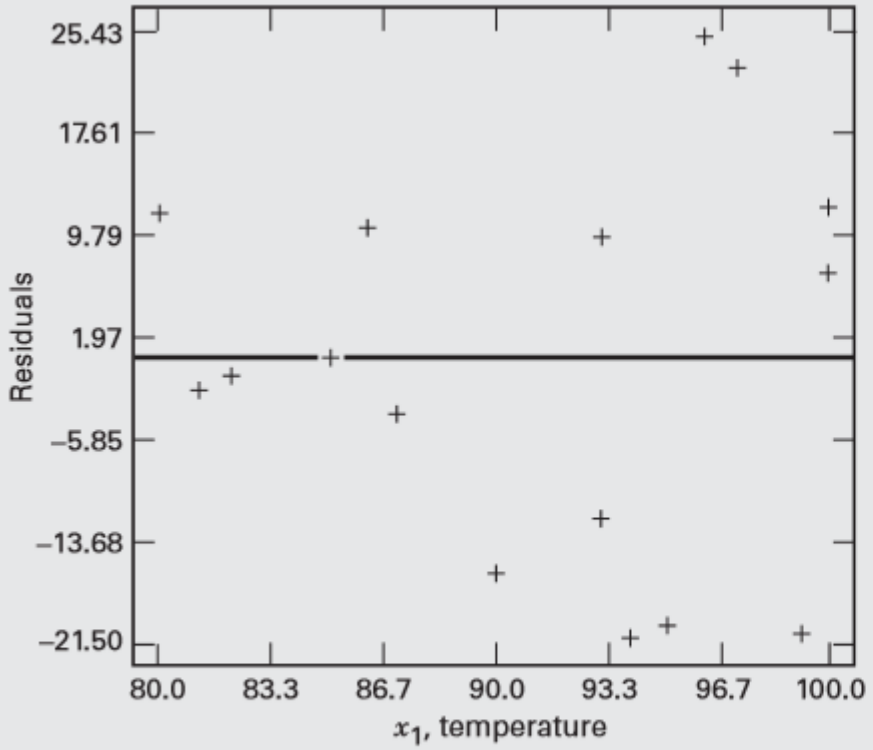
Observation i	y_i	Predicted Value $\hat{y}_i$	Residual e_i	h_{ii}	Studentized Residual	D_i	R-Student
1	2256	2244.5	11.5	0.350	0.87	0.137	0.87
2	2340	2352.1	−12.1	0.102	−0.78	0.023	−0.77
3	2426	2414.1	11.9	0.177	0.80	0.046	0.79
4	2293	2294.0	−1.0	0.251	−0.07	0.001	−0.07
5	2330	2346.4	−16.4	0.077	−1.05	0.030	−1.05
6	2368	2389.3	−21.3	0.265	−1.52	0.277	−1.61
7	2250	2252.1	−2.1	0.319	−0.15	0.004	−0.15
8	2409	2383.6	25.4	0.098	1.64	0.097	1.76
9	2364	2385.5	−21.5	0.142	−1.42	0.111	−1.48
10	2379	2369.3	9.7	0.080	0.62	0.011	0.60
11	2440	2416.9	23.1	0.278	1.66	0.354	1.80
12	2364	2384.5	−20.5	0.096	−1.32	0.062	−1.36
13	2404	2396.9	7.1	0.289	0.52	0.036	0.50
14	2317	2316.9	0.1	0.185	0.01	0.000	<0.01
15	2309	2298.8	10.2	0.134	0.67	0.023	0.66
16	2328	2332.1	−4.1	0.156	−0.28	0.005	−0.27



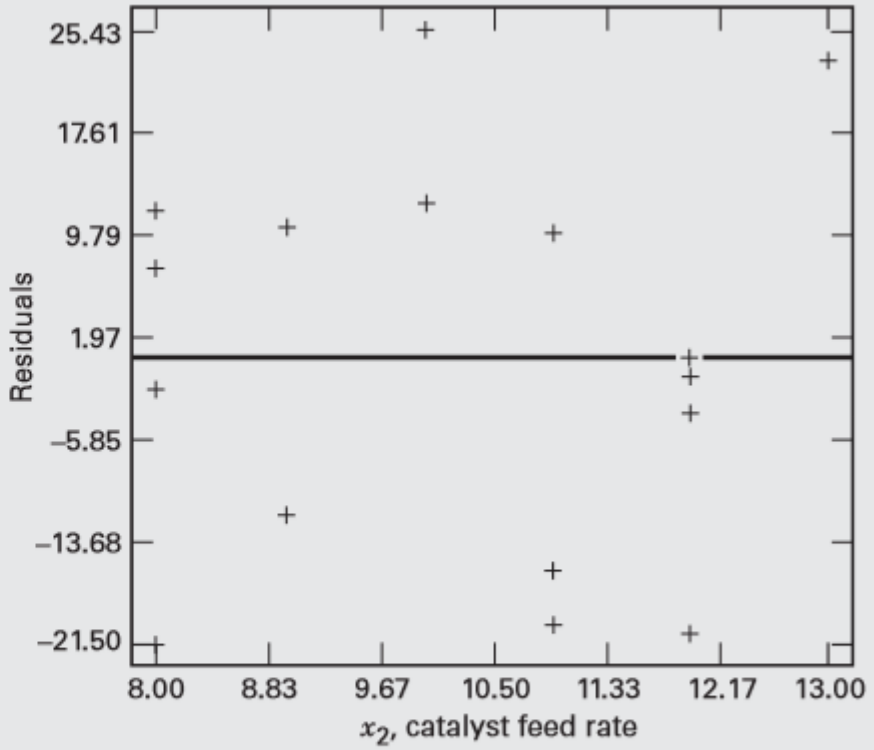
■ **FIGURE 10.1** **Normal probability plot of residuals, Example 10.1**



■ **FIGURE 10.2** **Plot of residuals versus predicted viscosity, Example 10.1**



■ **FIGURE 10.3** Plot of residuals versus x_1 (temperature), Example 10.1



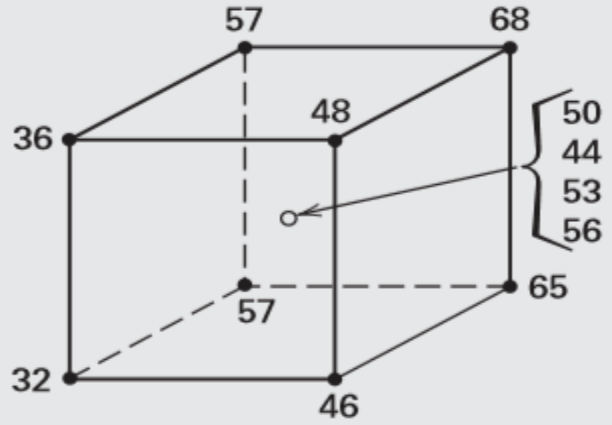
■ **FIGURE 10.4** Plot of residuals versus x_2 (feed rate), Example 10.1

EXAMPLE 10.2 Regression Analysis of a 2^3 Factorial Design

A chemical engineer is investigating the yield of a process. Three process variables are of interest: temperature, pressure, and catalyst concentration. Each variable can be run at a low and a high level, and the engineer decides to run a 2^3 design with four center points. The design and the resulting yields are shown in Figure 10.5, where we have shown both the natural levels of the design factor and the +1, −1 coded variable notation normally employed in 2^k factorial designs to represent the factor levels.

Process Variables				Coded Variables			Yield, y
Run	Temp (°C)	Pressure (psig)	Conc (g/l)	x_1	x_2	x_3	
1	120	40	15	−1	−1	−1	32
2	160	80	15	1	−1	−1	46
3	120	40	15	−1	1	−1	57
4	160	80	15	1	1	−1	65
5	120	40	30	−1	−1	1	36
6	160	80	30	1	−1	1	48
7	120	40	30	−1	1	1	57
8	160	80	30	1	1	1	68
9	140	60	22.5	0	0	0	50
10	140	60	22.5	0	0	0	44
11	140	60	22.5	0	0	0	53
12	140	60	22.5	0	0	0	56

$$x_1 = \frac{\text{Temp} - 140}{20}, \quad x_2 = \frac{\text{Pressure} - 60}{20}, \quad x_3 = \frac{\text{Conc} - 22.5}{7.5}$$



■ **FIGURE 10.5** Experimental design for Example 10.2

Suppose that the engineer decides to fit a main effects only model, say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

For this model the **X** matrix and **y** vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 68 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

The 2^3 is an orthogonal design, and even with the added center runs it is still orthogonal. Therefore

The 2^3 is an orthogonal design, and even with the added center runs it is still orthogonal. Therefore

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 612 \\ 45 \\ 85 \\ 9 \end{bmatrix}$$

Because the design is orthogonal, the $\mathbf{X}'\mathbf{X}$ matrix is *diagonal*, the required inverse is also diagonal, and the vector of least squares estimates of the regression coefficients is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1/12 & 0 & 0 & 0 \\ 0 & 1/8 & 0 & 0 \\ 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 1/8 \end{bmatrix} \begin{bmatrix} 612 \\ 45 \\ 85 \\ 9 \end{bmatrix}$$

$$= \begin{bmatrix} 51.000 \\ 5.625 \\ 10.625 \\ 1.125 \end{bmatrix}$$

The fitted regression model is

$$\hat{y} = 51.000 + 5.625x_1 + 10.625x_2 + 1.125x_3$$

As we have made use of on many occasions, the regression coefficients are closely connected to the effect estimates that would be obtained from the usual analysis of a 2^3 design. For example, the effect of temperature is (refer to Figure 10.5)

$$T = \bar{y}_{T^+} - \bar{y}_{T^-}$$
$$= 56.75 - 45.50 = 11.25$$

Notice that the regression coefficient for x_1 is

$$(11.25)/2 = 5.625$$

That is, the regression coefficient is exactly one-half the usual effect estimate. This will always be true for a 2^k design. As noted above, we used this result in Chapters 6 through 8 to produce regression models, fitted values, and residuals for several two-level experiments. This example demonstrates that the effect estimates from a 2^k design are least squares estimates.

The variance of the regression model parameter are found from the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. That is,

$$V(\hat{\beta}_0) = \frac{\sigma^2}{12}, \quad \text{and} \quad V(\hat{\beta}_i) = \frac{\sigma^2}{8}, i = 1,2,3.$$

The **relative variance** are

$$\frac{V(\hat{\beta}_0)}{\sigma^2} = \frac{1}{12} \quad \text{and} \quad \frac{V(\hat{\beta}_i)}{\sigma^2} = \frac{1}{8}, i = 1,2,3.$$

■ **TABLE 10.4**

Minitab Output for the Viscosity Regression Model, Example 10.1

Regression Analysis

The regression equation is

Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36 R-Sq = 92.7% R-Sq (adj) = 91.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

EXAMPLE 10.3 A 2^3 Factorial Design with a Missing Observation

Consider the 2^3 factorial design with four center points from Example 10.2. Suppose that when this experiment was performed, the run with all variables at the high level (run 8 in Figure 10.5) was missing. This can happen for a variety of reasons; the measurement system can produce a faulty reading, the combination of factor levels may prove infeasible, the experimental unit may be damaged, and so forth.

We will fit the main effects model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

using the 11 remaining observations. The $\mathbf{X}$ matrix and $\mathbf{y}$ vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

To estimate the model parameters, we form

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 11 & -1 & -1 & -1 \\ -1 & 7 & -1 & -1 \\ -1 & -1 & 7 & -1 \\ -1 & -1 & -1 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 544 \\ -23 \\ 17 \\ -59 \end{bmatrix}$$

Because there is a missing observation, the design is no longer orthogonal. Now

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \begin{bmatrix} 9.61538 \times 10^{-2} & 1.92307 \times 10^{-2} \\ 1.92307 \times 10^{-2} & 0.15385 \\ 1.92307 \times 10^{-2} & 2.88462 \times 10^{-2} \\ 1.92307 \times 10^{-2} & 2.88462 \times 10^{-2} \end{bmatrix}$$

$$\begin{bmatrix} 1.92307 \times 10^{-2} & 1.92307 \times 10^{-2} \\ 2.88462 \times 10^{-2} & 2.88462 \times 10^{-2} \\ 0.15385 & 2.88462 \times 10^{-2} \\ 2.88462 \times 10^{-2} & 0.15385 \end{bmatrix} \begin{bmatrix} 544 \\ -23 \\ 17 \\ -59 \end{bmatrix}$$

$$= \begin{bmatrix} 51.25 \\ 5.75 \\ 10.75 \\ 1.25 \end{bmatrix}$$

Therefore, the fitted model is

$$\hat{y} = 51.25 + 5.75x_1 + 10.75x_2 + 1.25x_3$$

Compare this model to the one obtained in Example 10.2, where all 12 observations were used. The regression coefficients are very similar. Because the regression coefficients are closely related to the factor effects, our conclusions would not be seriously affected by the missing observation. However, notice that the effect estimates are no longer orthogonal because $\mathbf{X}'\mathbf{X}$ and its inverse are no longer diagonal. Further more the variances of the regression coefficients are larger than they were in the original orthogonal design with no missing data.

When running a designed experiment, it is sometimes difficult to reach and hold the precise factor levels required by the design. Small discrepancies are not important, but large ones are potentially of more concern. Regression methods are useful in the analysis of a designed experiment where the experimenter has been unable to obtain the required factor levels.

To illustrate, the experiment presented in Table 10.5 shows a variation of the 2^3 design from Example 10.2, where many of the test combinations are not exactly the ones specified in the design. Most of the difficulty seems to have occurred with the temperature variable.

EXAMPLE 10.4 Inaccurate Levels in Design Factors

■ **TABLE 10.5**
Experimental Design for Example 10.4

	Process Variables			Coded Variables			
Run	Temp (°C)	Pressure (psig)	Conc (g/l)	x_1	x_2	x_3	Yield y
1	125	41	14	−0.75	−0.95	−1.133	32
2	158	40	15	0.90	−1	−1	46
3	121	82	15	−0.95	1.1	−1	57
4	160	80	15	1	1	−1	65
5	118	39	33	−1.10	−1.05	1.14	36
6	163	40	30	1.15	−1	1	48
7	122	80	30	−0.90	1	1	57
8	165	83	30	1.25	1.15	1	68
9	140	60	22.5	0	0	0	50
10	140	60	22.5	0	0	0	44
11	140	60	22.5	0	0	0	53
12	140	60	22.5	0	0	0	56

We will fit the main effects model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

The **X** matrix and **y** vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -0.75 & -0.95 & -1.133 \\ 1 & 0.90 & -1 & -1 \\ 1 & -0.95 & 1.1 & -1 \\ 1 & 1 & 0 & -1 \\ 1 & -1.10 & -1.05 & 1.4 \\ 1 & 1.15 & -1 & 1 \\ 1 & -0.90 & 1 & 1 \\ 1 & 1.25 & 1.15 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 68 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

To estimate the model parameters, we need

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 0.60 & 0.25 & 0.2670 \\ 0.60 & 8.18 & 0.31 & -0.1403 \\ 0.25 & 0.31 & 8.5375 & -0.3437 \\ 0.2670 & -0.1403 & -0.3437 & 9.2437 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 612 \\ 77.55 \\ 100.7 \\ 19.144 \end{bmatrix}$$

Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \begin{bmatrix} 8.37447 \times 10^{-2} & -6.09871 \times 10^{-3} & -2.33542 \times 10^{-3} & -2.59833 \times 10^{-3} \\ -6.09871 \times 10^{-3} & 0.12289 & -4.20766 \times 10^{-3} & 1.88490 \times 10^{-3} \\ -2.33542 \times 10^{-3} & -4.20766 \times 10^{-3} & 0.11753 & 4.37851 \times 10^{-3} \\ -2.59833 \times 10^{-3} & 1.88490 \times 10^{-3} & 4.37851 \times 10^{-3} & 0.10845 \end{bmatrix} \begin{bmatrix} 612 \\ 77.55 \\ 100.7 \\ 19.144 \end{bmatrix} = \begin{bmatrix} 50.49391 \\ 5.40996 \\ 10.16316 \\ 1.07245 \end{bmatrix}$$

The fitted regression model, with the coefficients reported to two decimal places, is

$$\hat{y} = 50.49 + 5.41x_1 + 10.16x_2 + 1.07x_3$$

Comparing this to the original model in Example 10.2, where the factor levels were exactly those specified by the design, we note very little difference. The practical interpretation of the results of this experiment would not be seriously affected by the inability of the experimenter to achieve the desired factor levels exactly.

EXAMPLE 10.5 De-aliasing Interactions in a Fractional Factorial

We observed in Chapter 8 that it is possible to de-alias interactions in a fractional factorial design by a process called fold over. For a resolution III design, a full fold over is constructed by running a second fraction in which the signs are reversed from those in the original fraction. Then the combined design can be used to de-alias all main effects from the two-factor interactions.

A difficulty with a full fold over is that it requires a second group of runs of identical size as the original design. It is usually possible to de-alias certain interactions of interest by augmenting the original design with fewer runs than required in a full fold over. The partial fold-over technique was used to solve this problem. Regression methods are an easy way to see how the partial fold-over technique works and, in some cases, find even more efficient fold-over designs.

To illustrate, suppose that we have run a 2_{IV}^{4-1} design. Table 8.3 shows the principal fraction of this design, in which $I = ABCD$. Suppose that after the data from the first eight trials were observed, the largest effects were A, B, C, D (we ignore the three-factor interactions that are aliased with these main effects) and the $AB + CD$ alias chain. The other two alias chains can be ignored, but clearly either AB, CD, or both two-factor interactions are large. To find out which interactions are important, we could, of course, run the alternate fraction, which would require another eight trials. Then all 16 runs could be used to estimate the main effects and the two-factor interactions. An alternative would be to use a partial fold over involving four additional runs.

It is possible to de-alias AB and CD in fewer than four additional trials. Suppose that we wish to fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{34} x_3 x_4 + \epsilon$$

where x_1, x_2, x_3, and x_4 are the coded variables representing A, B, C, and D. Using the design in Table 8.3, the $\mathbf{X}$ matrix for this model is

$$\mathbf{X} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

where we have written the variables above the columns to facilitate understanding. Notice that the x_1x_2 column is identical to the x_3x_4 column (as anticipated, because AB or x_1x_2 is aliased with CD or x_3x_4), implying a linear dependency in the columns of $\mathbf{X}$. Therefore, we cannot estimate both β_{12} and β_{34} in the model. However, suppose that we add a single run $x_1 = -1$, $x_2 = -1$, $x_3 = -1$, and $x_4 = 1$ from the alternate fraction to the original eight runs.

The **X** matrix for the model now becomes

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 \end{array} \\ \begin{bmatrix} -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix} \end{array}$$

Notice that the columns x_1x_2 and x_3x_4 are now no longer identical, and we can fit the model including both the x_1x_2 (AB) and x_3x_4 (CD) interactions. The magnitudes of the regression coefficients will give insight regarding which interactions are important.

Although adding a single run will de-alias the AB and CD interactions, this approach does have a disadvantage. Suppose that there is a time effect (or a block effect) between the first eight runs and the last run added above. Add a column to the **X** matrix for blocks, and you obtain the following:

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 & \text{block} \end{array} \\ \begin{bmatrix} -1 & -1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & -1 & 1 \end{bmatrix} \end{array}$$

We have assumed the block factor was at the low or " − " level during the first eight runs, and at the high or " + " level during the ninth run. It is easy to see that the sum of the cross products of every column with the block column does not sum to zero, meaning that blocks are no longer orthogonal to treatments, or that the block effect now affects the estimates of the model regression coefficients. To block orthogonally, you must add an even number of runs. For example, the four runs

x_1	x_2	x_3	x_4
−1	−1	−1	1
1	−1	−1	−1
−1	1	1	1
1	1	1	−1

will de-alias *AB* from *CD* and allow orthogonal blocking (you can see this by writing out the **X** matrix as we did previously). This is equivalent to a partial fold over, in terms of the number of runs that are required.

Hypothesis Testing in Multiple Regression: Test for Significance of Regression

The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad \textbf{(10.20)}$$

$$H_1: \beta_j \neq 0 \quad \text{for at least one } j$$

Rejection of H_0 in Equation 10.20 implies that at least one of the regressor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model. The test procedure involves an analysis of variance partitioning of the total sum of squares SS_T into a sum of squares due to the model (or to regression) and a sum of squares due to residual (or error), say

$$SS_T = SS_R + SS_E \qquad \textbf{(10.21)}$$

Now if the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true, then SS_R/σ^2 is distributed as χ^2_k, where the number of degrees of freedom for χ^2 is equal to the number of regressor variables in the model. Also, we can show that SS_E/σ^2 is distributed as χ^2_{n-k-1} and that SS_E and SS_R are independent. The test procedure for $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is to compute

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E} \tag{10.22}$$

and to reject H_0 if F_0 exceeds $F_{\alpha,k,n-k-1}$. Alternatively, we could use the P-value approach to hypothesis testing and, reject H_0 if the P-value for the statistic F_0 is less than α. The test is usually summarized in an analysis of variance table such as Table 10.6.

■ **TABLE 10.6**

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

the regression sum of squares is

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{10.23}$$

and the error sum of squares is

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{10.24}$$

and the total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{10.25}$$

- Table 10.4 (Present in slide 18) shows some of the output from Minitab for the viscosity regression model in Example 10.1.
- The upper portion in this display is the analysis of variance for the model. The test of significance of regression in this example involves the hypotheses:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_j \neq 0 \qquad \text{for at least one } j$$

The P-value in Table 10.4 for the F statistic (Equation 10.22) is very small, so we would conclude that at least one of the two variables—temperature (x_1) and feed rate (x_2)—has a nonzero regression coefficient.

R^2

Table 10.4 also reports the coefficient of multiple determination R^2, where

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{10.26}$$

Just as in designed experiments, R^2 is a measure of the amount of reduction in the variability of y obtained by using the regressor variables $x_1, x_2, \ldots, x_k$ in the model. However, as we have noted previously, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a variable to the model will always increase R^2, regardless of whether the additional variable is statistically significant or not. Thus, it is possible for models that have large values of R^2 to yield poor predictions of new observations or estimates of the mean response.

Because R^2 always increases as we add terms to the model, some regression model builders prefer to use an **adjusted R^2 statistic** defined as

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2) \qquad \textbf{(10.27)}$$

In general, the adjusted R^2 statistic will not always increase as variables are added to the model. In fact, if unnecessary terms are added, the value of R^2_{adj} will often decrease.

Tests on Individual Regression Coefficients and Groups of Coefficients

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.

The hypotheses for testing the significance of any individual regression coefficient, say β_j, are

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

If $H_0: \beta_j = 0$ is not rejected, then this indicates that x_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{10.28}$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. The null hypothesis $H_0: \beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. Note that this is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all the other regressor variables x_i $(i \neq j)$ that are in the model.

The denominator of Equation 10.28, $\sqrt{\hat{\sigma}^2 C_{jj}}$, is often called the **standard error** of the regression coefficient $\hat{\beta}_j$. That is,

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{10.29}$$

Therefore, an equivalent way to write the test statistic in Equation (10.28) is

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{10.30}$$

- Consider the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \qquad (10.32)$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$.

$SS_R(\boldsymbol{\beta})$ is called the regression sum of squares due to $\boldsymbol{\beta}$.

- It can be mathematically expressed as:

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p \text{ degrees of freedom})$$

The regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) \tag{10.35}$$

This sum of squares has r degrees of freedom. It is the "extra sum of squares" due to $\boldsymbol{\beta}_1$. Note that $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to inclusion of variables $x_1, x_2, \ldots, x_r$ in the model.

Now, $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is independent of MS_E, and the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{Ss_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{MS_E} \tag{10.36}$$

If $F_0 > F_{\alpha,r,n-p}$, we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_1$ is not zero, and, consequently, at least one of the variables $x_1, x_2, \ldots, x_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. Some authors call the test in Equation 10.36 a **partial F test**.

EXAMPLE 10.6

Consider the viscosity data in Example 10.1. Suppose that we wish to investigate the contribution of the variable x_2 (feed rate) to the model. That is, the hypotheses we wish to test are

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

This will require the extra sum of squares due to β_2, or

$$SS_R(\beta_2|\beta_1, \beta_0) = SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0, \beta_1)$$

$$= Ss_R(\beta_1, \beta_2|\beta_0) - SS_R(\beta_2|\beta_0)$$

Now from Table 10.4, where we tested for significance of regression, we have

$$SS_R(\beta_1, \beta_2|\beta_0) = 44{,}157.1$$

which was called the model sum of squares in the table. This sum of squares has two degrees of freedom.

The reduced model is

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

The least squares fit for this model is

$$\hat{y} = 1652.3955 + 7.6397x_1$$

and the regression sum of squares for this model (with one degree of freedom) is

$$SS_R(\beta_1|\beta_0) = 40{,}840.8$$

Note that $SS_R(\beta_1|\beta_0)$ is shown at the bottom of the Minitab output in Table 10.4 under the heading "Seq SS."

Therefore,

$$SS_R(\beta_2|\beta_0, \beta_1) = 44{,}157.1 - 40{,}840.8$$

$$= 3316.3$$

with $2 - 1 = 1$ degree of freedom. This is the increase in the regression sum of squares that results from adding x_2 to a model already containing x_1, and it is shown at the bottom of the Minitab output on Table 10.4. To test $H_0: \beta_2 = 0$, from the test statistic we obtain

$$F_0 = \frac{Ss_R(\beta_2|\beta_0, \beta_1)/1}{MS_E} = \frac{3316.3/1}{267.604} = 12.3926$$

Note that MS_E from the full model (Table 10.4) is used in the denominator of F_0. Now, because $F_{0.05,1,13} = 4.67$, we would reject $H_0: \beta_2 = 0$ and conclude that x_2 (feed rate) contributes significantly to the model.

Because this partial F test involves only a single regressor, it is equivalent to the t-test because the square

of a t random variable with v degrees of freedom is an F random variable with 1 and v degrees of freedom. To see this, note from Table 10.4 that the t-statistic for $H_0: \beta_2 = 0$ resulted in $t_0 = 3.5203$ and that $t_0^2 = (3.5203)^2 = 12.3925 \simeq F_0$.

Confidence Intervals on the Individual Regression Coefficients

Because the least squares estimator $\hat{\boldsymbol{\beta}}$ is a linear combination of the observations, it follows that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Then each of the statistics

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0, 1, \ldots, k \tag{10.37}$$

is distributed as t with $n - p$ degrees of freedom, where C_{jj} is the (jj)th element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Equation 10.17. Therefore, a $100(1 - \alpha)$ percent confidence interval for the regression coefficient β_j, $j = 0, 1, \ldots, k$, is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{10.38}$$

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} \tag{10.17}$$

EXAMPLE 10.7

We will construct a 95 percent confidence interval for the parameter β_1 in Example 10.1. Now $\hat{\beta}_1 = 7.62129$, and because $\hat{\sigma}^2 = 267.604$ and $C_{11} = 1.429184 \times 10^{-3}$, we find that

$$\hat{\beta}_1 - t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$7.62129 - 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})} \le \beta_1 \le 7.62129 + 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})}$$

$$7.62129 - 2.16(0.6184) \le \beta_1 \le 7.62129 + 2.16(0.6184)$$

and the 95 percent confidence interval on β_1 is

$$6.2855 \le \beta_1 \le 8.9570$$

Prediction of New Response Observations

A regression model can be used to predict future observations on the response y corresponding to particular values of the regressor variables, say $x_{01}, x_{02}, \ldots, x_{0k}$. If $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, then a point estimate for the future observation y_0 at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is computed from Equation 10.39:

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

A $100(1 - \alpha)$ **percent prediction interval** for this future observation is

$$\hat{y}(\mathbf{x}_0) = t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0$$
$$\le \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \quad \textbf{(10.42)}$$

In predicting new observations and in estimating the mean response at a given point $x_{01}, x_{02}, \ldots, x_{0k}$, we must be careful about extrapolating beyond the region containing the original observations. It is very possible that a model that fits well in the region of the original data will no longer fit well outside of that region.

- The prediction interval in Equation 10.42 has many useful applications.
- One of these is in confirmation experiments following a factorial or fractional factorial experiment. In a confirmation experiment, we are usually testing the model developed from the original experiment to determine if our interpretation was correct.
- Often we will do this by using the model to predict the response at some point of interest in the design space and then comparing the predicted response with an actual observation obtained by conducting another trial at that point.
- A useful measure of confirmation is to see if the new observation falls inside the prediction interval on the response at that point.